ESD-TR-67-457

CONFIDENCE IN RECALL IN PAIRED-ASSOCIATE LEARNING EXPERIMENTS

ELIZABETH H. NICOL
THORTON B. ROBY
FRANCIS M. FARRELL

MAY 1967

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts, 01730

LEGAL NOTICE

When U. S. Government drawings, specifications or other data are used for any purpose other than a definitely related government procurement operation, the government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

OTHER NOTICES

Do not return this copy. Retain or destroy.

ESD-TR-67-457

CONFIDENCE IN RECALL IN PAIRED-ASSOCIATE LEARNING EXPERIMENTS

ELIZABETH H. NICOL
THORTON B. ROBY
FRANCIS M. FARRELL

MAY 1967

DECISION SCIENCES LABORATORY
ELECTRONIC SYSTEMS DIVISION
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE
L. G. Hanscom Field, Bedford, Massachusetts, 01730

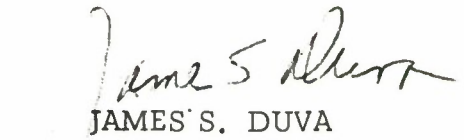
FOREWORD

The experimental series reported here were part of the in-house research program of the Decision Sciences Laboratory. They originated under Task 96782 of Project 9678 entitled "Computer-Man Relationships in Information Processing Command and Control Systems." When that project was terminated, the research effort continued under Project 7682, "Man-Computer Information Processing." Task 768203, which supported the research, was entitled "Decision-Making in Computer Environments" and was concerned with seeking techniques for improving the quality of decisions reached in man-machine information processing systems.

Acknowledgment is made of the assistance of Mr. Arthur Marcus, who conducted the testing in Series 1, and to Captain Richard S. Gibson, who played a major role in the design, administration and evaluation of Series 3.

This technical report has been reviewed and is approved.

WILLIAM H. SUMBY
Project Officer
Decision Sciences Laboratory



JAMES S. DUVA
Technical Director
Decision Sciences Laboratory

ABSTRACT

The problem of estimating the accuracy of ones own recollections was investigated in four experiments under a variety of conditions. Subjects were shown a series of paired words; then they were shown the first member of the pairs and asked to recall the second member of each. Along with each attempt at recall the subjects were asked to give a confidence rating on a scale from 1 to 5. In all, 180 subjects were tested for a total of 11,200 trials. The confidence results are highly significant, indicating that subjects were able to discriminate their correct recollections from mere guesses. Comparisons are presented showing how realism of confidence varies over the main experimental treatments: variations in meaningfulness of material, one versus two exposures to paired-associate lists, and presence of varying amounts of irrelevant material in the acquisition lists.

TABLE OF CONTENTS

CONFIDENCE IN RECALL

IN PAIRED-ASSOCIATE LEARNING EXPERIMENTS

Introduction

This research is concerned with the problem of whether subjects can give useful estimates of the accuracy of their own recall for different kinds of material learned under different conditions. Psychologists have long been interested in how accurately human beings can evaluate their own performance levels. For at least eighty years experimenters have been using measures of confidence or subjective probability along with various experimental tasks. The general finding is that a positive monotonic relationship exists between confidence and some measure of performance, showing that subjects are in some sense realistic about the adequacy of their performance. This relation has appeared over a wide variety of tasks: true-false tests of factual material, learning situations, psychophysical judgments and so on.

In some of the psychophysical work (e.g., Pollack and Decker, 1958; Carterette and Cole, 1959) the confidence-accuracy relationship has appeared invariant -- that is, the form of the confidence-accuracy relation does not vary with the difficulty level of the judgmental task. Later work with filtered speech (Decker and Pollack, 1958) suggests that this finding may

not be as general as was originally believed for the psychophysical situation.

Nickerson and McGoldrick (1963) of this laboratory have been concerned with the confidence-correctness relationship in non-psychophysical judgmental tasks. They developed tests of varying difficulty involving comparisons of sizes of states of the union; the unique feature of these tests is that an objective criterion of difficulty was used in the test construction. Nickerson and McGoldrick found that the confidence-correctness relation did vary with task difficulty. They concluded that "the value of confidence expressions as indicators of the probable correctness of 'intellectual' judgment varies considerably with the difficulty of the judgments involved" (Nickerson and McGoldrick, 1963).

Our own work in this area began back in 1960 when we undertook a series of studies of short-term memory. These experiments deal with the use of a confidence rating by subjects to indicate the likelihood that their attempted recall of paired associates is in fact correct. As with many confidence-rating studies, this aspect was a somewhat secondary adjunct to research on another topic; in this case, the major project was concerned with some of the variables influencing the information-processing capacities of the individual. Because the information-processing project was our primary concern, the major experimental treatments and comparisons were dictated by the objectives of that project. Since this phase of the work has been reported elsewhere (Nicol, Farrell, and Roby, 1962), the details will be mentioned here only insofar as they are directly relevant to the confidence study.

The Experiments

Four experimental series were completed in all. All used a modified paired-associate learning situation: a number of nonsense syllables were associated with a relatively small number of categories. In some sections 32 nonsense syllables were associated with 4 categories: in other sections 32 syllables were associated

with 8 categories. In other words, in the 4-category sections, eight different nonsense syllables were paired at some time or other with one given category name, such as the color 'red' and so on. In the 8-category sections, each category name was paired at some time in the acquisition list with each of four different nonsense syllables. The basic condition, then, is that 32 nonsense syllables were distributed over either four or eight category names. As will be seen later, one of the principal questions was whether it is easier to learn the relation of 32 syllables to four categories than to eight.

Another main variable in the acquisition lists was the degree of meaningfulness or familiarity of the response categories. This has been shown (Underwood and Schulz, 1960) to be of more importance for retention than is the meaningfulness or association value of the stimuli - in this case the nonsense syllables. In two of our experimental series, names of colors were used for the high-association value categories, while names of trees were used for low-association value; according to the Thorndike-Lorge word frequency count, the color names are responded to much more frequently in everyday experience than are the names of trees. In a third experiment, in an attempt to increase the contrast in meaningfulness or familiarity, tree names as categories were compared with nonsense syllables themselves used as categories in association with the basic 32 nonsense syllables. For these nonsense syllable categories, we selected eight syllables from Hull's lowest association-value range (Stevens, 1951) and employed these as the categories with which the customary 32 nonsense syllables were associated.

Our fourth experiment did not make the meaningfulness comparison; in this, the primary objective was a comparison of retention after one presentation of the acquisition lists with that after two presentations. For this purpose, only the names of colors were used as categories associated with the 32 nonsense syllables.

In three of the series, we studied the effects of the presence of irrelevant material on retention. The irrelevant material consisted of interspersed nonsense syllables that were never associated with a category. In two series these unassociated syllables were as numerous as were the paired-associates. In the third series, five different levels of 'density' of such 'scrap' material were sampled.

Table 1 summarizes the major features of each of the four series.

In all cases the procedure was the same: the acquisition series of paired-associates (with or without irrelevant material) was presented on slides by means of timed projection. They were shown to groups of 4 to 8 subjects at once. Immediately after the acquisition trials came the retention test in which the nonsense syllables appeared successively in query form on the projection screen; the subjects were asked to write down the proper category name. On the record form the space for each response was followed by the numbers 1 to 5. Subjects were asked to circle '5' if they were positive of their answer, to circle '1' if they were sure the answer was merely a guess, and to use the remaining numbers for intermediate degrees of certainty. The measure of recall was the number of syllables correctly placed with categories. Subjects received no feedback at any time during or after the experiments.

Results and Discussion

Because of the variety of comparisons, the presentation of results will be streamlined to the extent that graphs will be relied on heavily and the citing of

TABLE 1

Basic Details of Experimental Series

Data	Number and Type of Subjects	Trials per Subject	Total Trials per Series*	Classes of Associated Categories	No. of Categories	Principal Objectives
SERIES 1	16 airmen	64	1024 (995)	Colors, Trees	8	Comparison of categories at two levels of familiarity (Colors vs. Trees).
SERIES 2	32 airmen	64	2048 (2032)	Colors, Trees	4, 8	Same as above. Comparison of 4 categories versus 8 categories. Effect of interspersing irrelevant nonsense syllables.
SERIES 3	72 college students	64	4608 (4538)	Colors	4, 8	Learning: retention tested after one and after two presentations of acquisition lists. Effect of interspersing irrelevant nonsense syllables: 5 'density' levels tested.
SERIES 4	60 college students	64	3776 (3637)	Trees, Nonsense Syllables	4, 8	Same as for Series 2 except for use of 'nonsyls' as categories to increase range in familiarity comparison
TOTAL:	180		11,456 (11,202)			

*Figures in parentheses indicate the number of trials usable in confidence study; occasionally subjects failed to give confidence rating along with category response.

statistics will be minimized.

The relation of confidence to accuracy was measured by chi-square tests of 2 x 5 contingency tables whose cells contained the number of right and number of wrong responses for each of the five confidence ratings. There may be some reservations about use of chi-square on data consisting of more than one observation per subject, but in all cases here the chi-squares range far beyond the scope of any table and are clearly significant.

Series 1

The relationship of response correctness to the confidence scale is shown for Series 1 in Figure 1. The confidence-correctness relation is highly significant with both colors and trees as category names. Since 8 categories were used throughout this series, a purely chance score here would be 1/8 or 12.5 per cent. Subjects' scores tend to approach the 'ideal' function most closely at the chance point, to deviate from it at mid-range confidence ratings and to return more toward the upper end of the scale.

It is axiomatic that an averaging procedure, such as that shown in Fig. 1, may in fact run the risk of misrepresenting the trends of individual subjects. In this particular area of confidence ratings, it is possible for significant results to be an artefact occasioned only by the pooling of high-scoring and low-scoring subjects. For all of our series we have made special studies of high-scoring and low-scoring subjects and their use of the confidence scales. Here we will give only the results of the first two series pooled. (See Figure 2.) The series are given pooled because in Series 1 alone there were too few chance-scoring subjects (only two) to afford a basis for comparison. Consequently, Series 1 subjects were pooled with those of the 8-category section of Series 2. The 4-category section of Series 2 is also shown in this graph. The consistently high success ratios for confidence rating of 5 and the general form of the curves suggest that the relationship in Figure 1 and later figures is a genuine one and is fairly representative of the general run of subject.

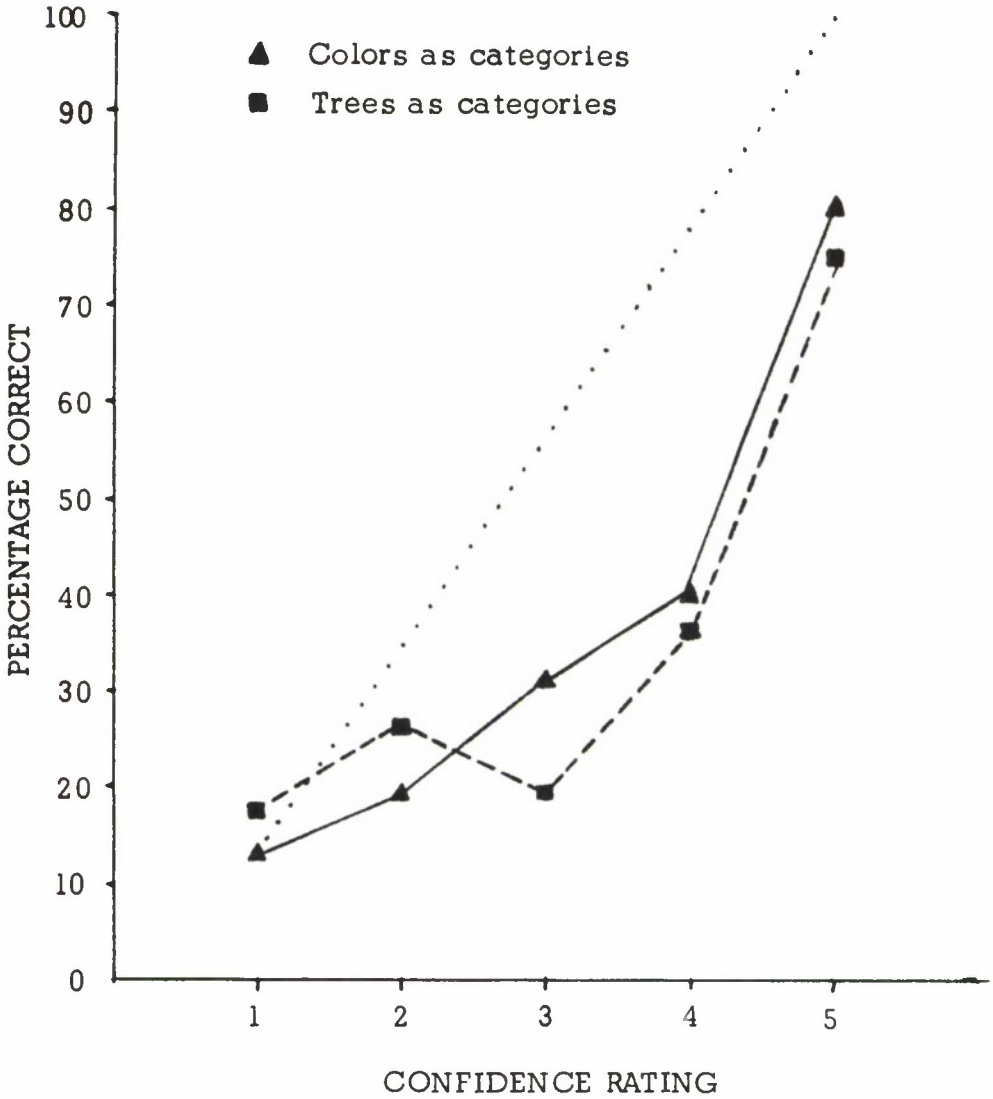


Fig. 1. Response correctness and confidence ratings for Series 1. The abscissa represents the subjects' confidence ratings, ranging from '1,' a mere guess, to '5,' complete certainty. The ordinate is the proportion of correct recalls over the total number of responses at each confidence category. The dotted line represents 'ideal' agreement between percentage correct and confidence level.

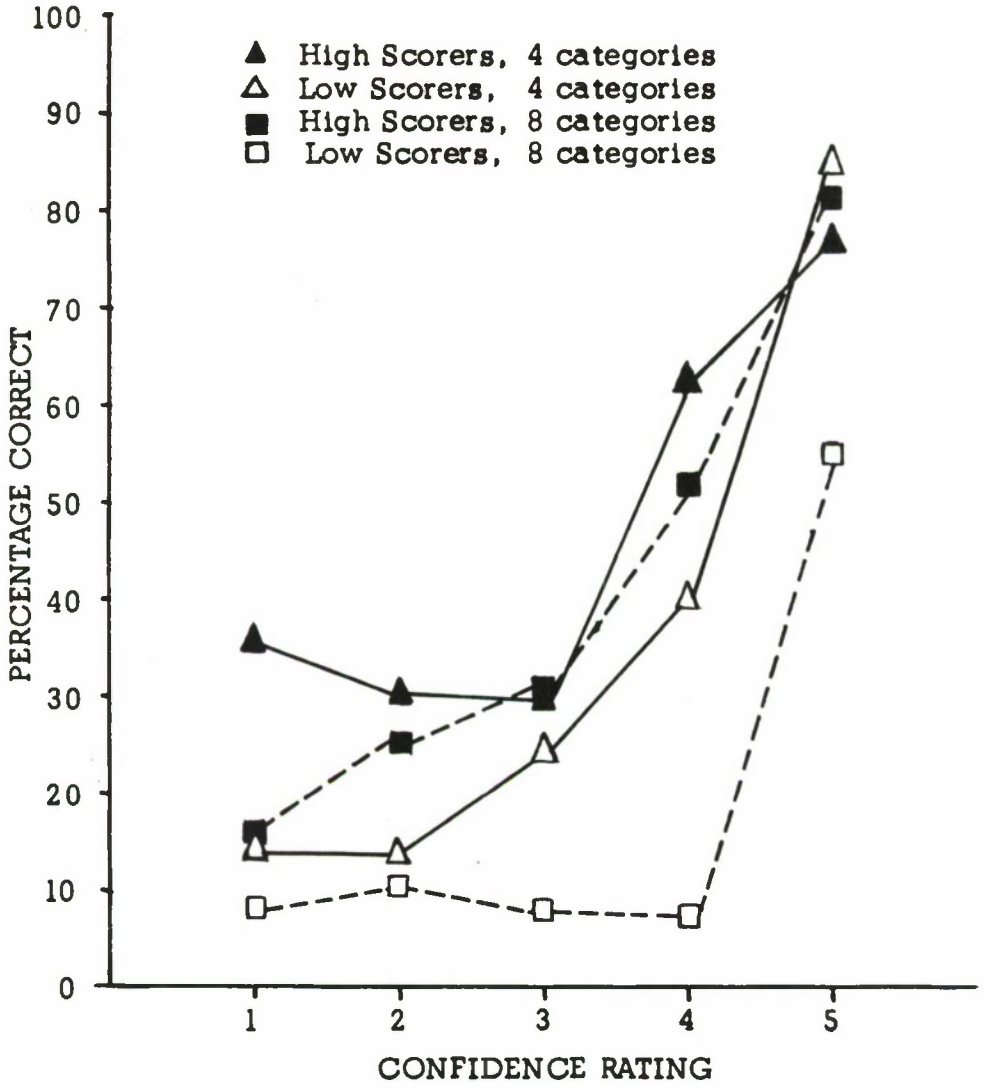


Fig. 2. Comparisons of high-scoring and low-scoring subjects in the 4-category and 8-category sections of Series 1 and 2, pooled. 'High Scorers' are those whose retention scores were in the top quartile of their series. 'Low Scorers' are those whose retention scores were at or below the chance level. For the 4-category sections the chance score is 25%; for the 8-category sections the chance score is 12.5%.

Similar results emerged from detailed analyses of the later series. Thus, for the remainder of this brief report we shall cite only pooled results.

Now let us return to Series 1 and our original question concerning effects of the degree of response meaningfulness or familiarity as represented in the colors - trees contrast. The previous report on this series (Nicol, Farrell and Roby, 1962) noted that colors were recalled significantly more often than were trees. Now, the confidence ratings reflect this preference. Considering only frequency of usage of the confidence ratings, we find that the mean confidence score for colors was 3.49 while that for trees was 3.19. This difference is significant with a probability of less than .01 as measured by the chi-square approximation to the Kolmogorov-Smirnov test of the difference between two distributions.

Series 2 and 4

These two series may be reported together since they used the same experimental design and stimulus-list formats. The differences were two: (1) airmen were subjects in Series 2, while college students served in Series 4; and (2) for the comparisons of degree of meaningfulness or familiarity, Series 2 used colors and trees as category names, while Series 4 used trees and low-association value nonsense syllables as category names. (See Table 1.)

Figures 3 and 4 present the results of these two series for the main experimental conditions. In all cases the accuracy of recall in relation to confidence rating is very highly significant.

It is apparent also that the results for colors and trees in Series 2 are closely parallel while the same is true for those for trees and nonsense syllables in Series 4.

As with Series 1, the confidence-correctness results come close to the chance line at the lower end of the scale, show the rather characteristic sag in the middle and approach the 'ideal' function line only at the upper end of the confidence scale.

The comparisons of 4 categories versus 8 categories showed no differences in

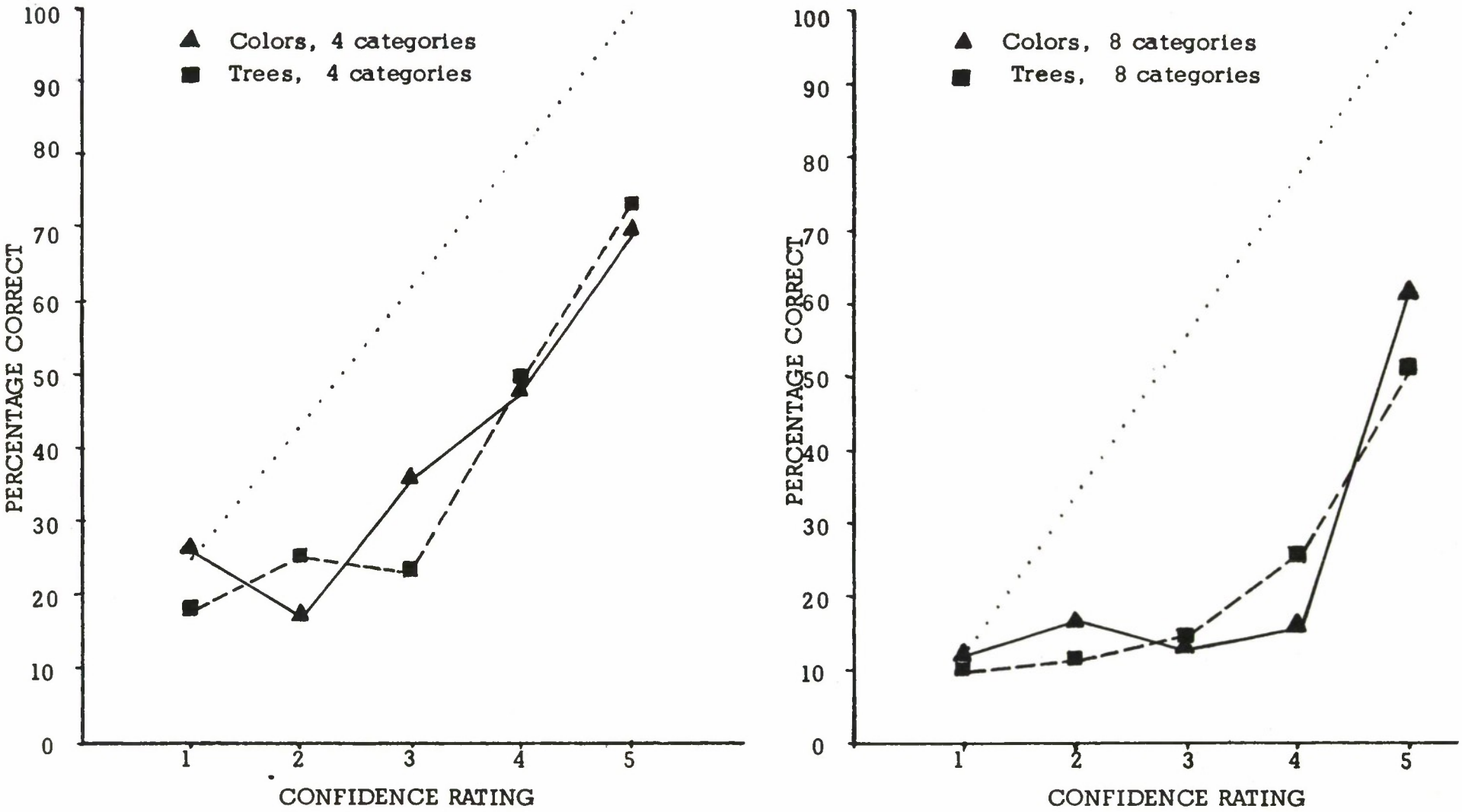


Fig. 3. Response correctness and confidence ratings for Series 2. On the left are the results for the 4-category section where the chance score is 25%. On the right are the results for the 8-category section where the chance score is 12.5%.

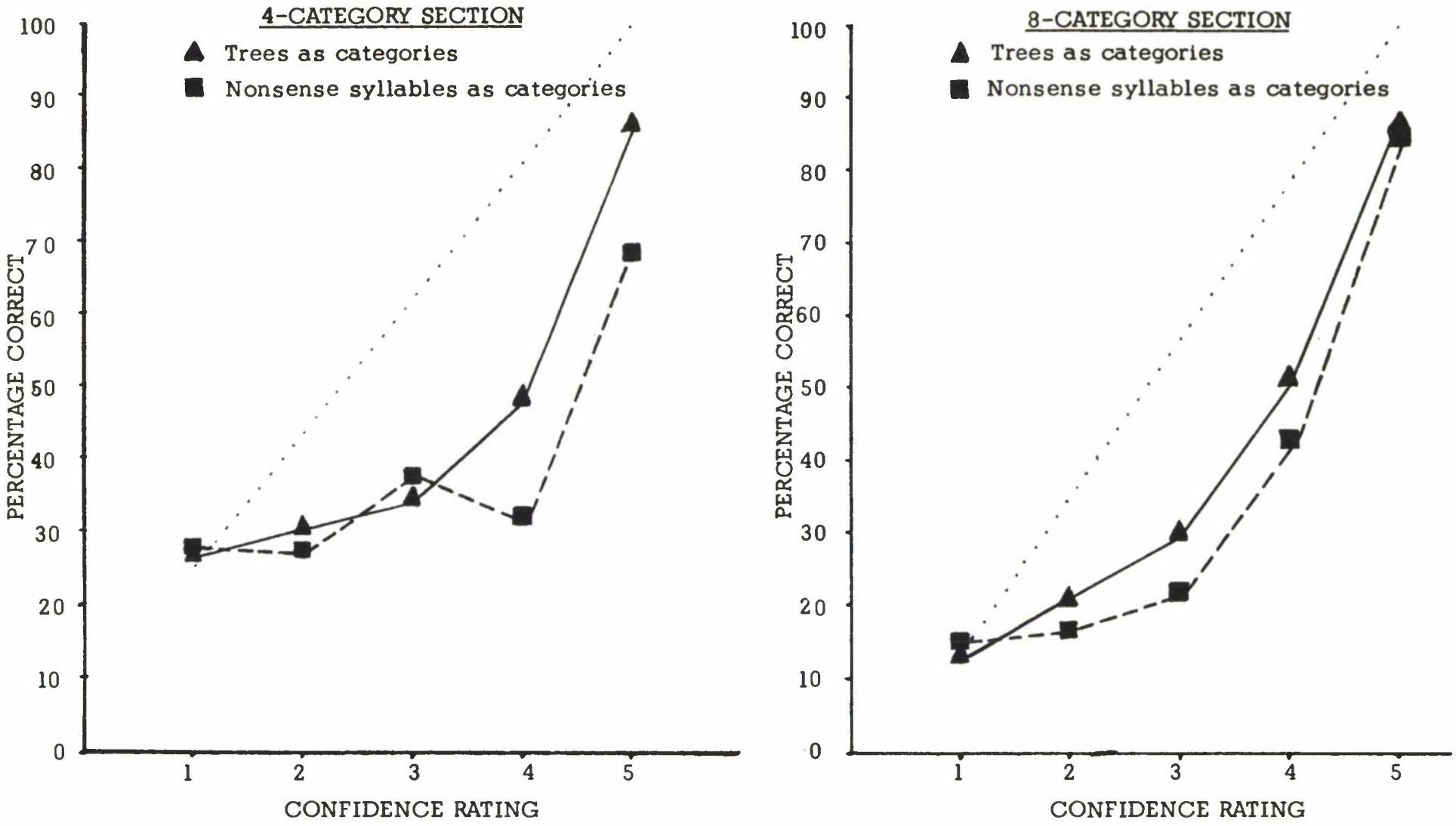


Fig. 4. Response correctness and confidence ratings for Series 4. On the left are the results for the 4-category section (chance score = 25%). On the right are the results for the 8-category section (chance score = 12.5%).

recall and no differences in mean confidence scores. This is true of both series.

The picture with regard to the effects of irrelevant material or 'noise' is not clear in relation to recall. In Series 2 the presence of irrelevant material was associated with a significant decrement in recall. In Series 4 there was no effect attributable to such 'noise.' (Incidentally, Series 3 which provided an elaborate test of varying degrees of 'noise' was also without significant trends on this point.)

For both Series 2 and 4, mean confidence scores for the 'noise' conditions were in general slightly higher than those for 'no-noise' and in Series 4 this curious reversal of expected trend approaches the .05 significance level. (See Figures 5 and 6.)

Series 3

If we turn to Series 3 where a more extensive study of noise effects was made, we find little light here either. In the four sets of data in Figure 7, there are no significant differences between retention scores for paired associates presented alone and retention scores for paired associates interspersed with irrelevant syllables. Neither do the mean confidence scores differ. Thus we can only say that we find no significant effects attributable to the influence of irrelevant material on the basic retention task.

The main point of Series 3, however, was to compare retention and confidence scores after one presentation of the acquisition list with similar scores after a second presentation. The procedure was the same as that for the other series except that here after the recall test, the entire procedure was immediately repeated.

The confidence-correctness relation for all results in Figure 7 is highly significant. As with the other experiments, the subjects' results approach the

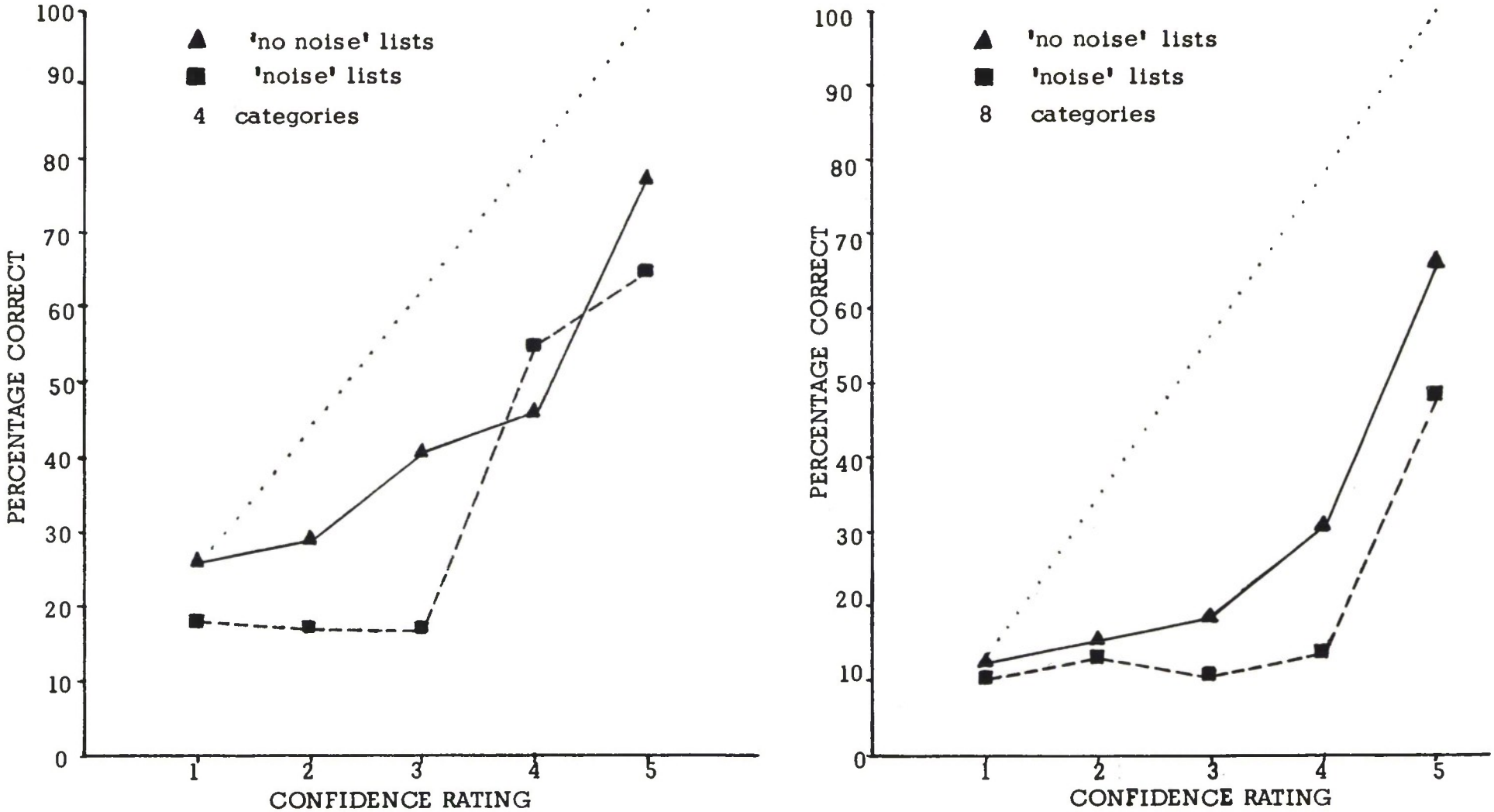


Fig. 5. For Series 2, the correctness-confidence relations are plotted for the paired-associate lists having irrelevant material interspersed ('noise') and for lists without such material ('no noise'). On the left are the results for the 4-category section (chance score=25%); on the right are the results for the 8-category section (chance score=12.5%).

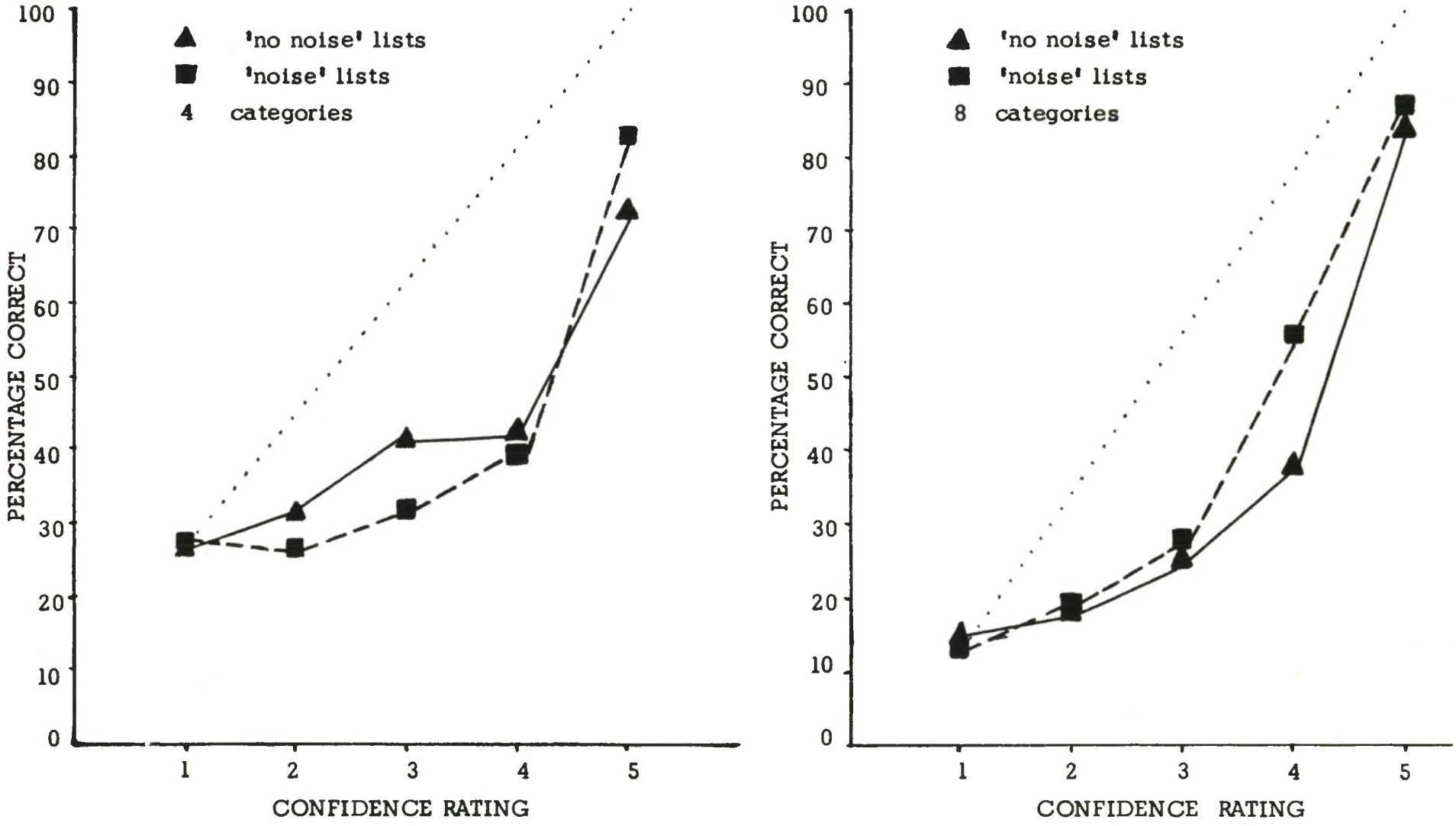


Fig. 6. For Series 4, the correctness-confidence relations are plotted for the paired-associate lists having irrelevant material interspersed ('noise') and for lists without such material ('no noise'). On the left are the results for the 4-category section (chance score = 25%); on the right are the results for the 8-category section (chance score = 12.5%).

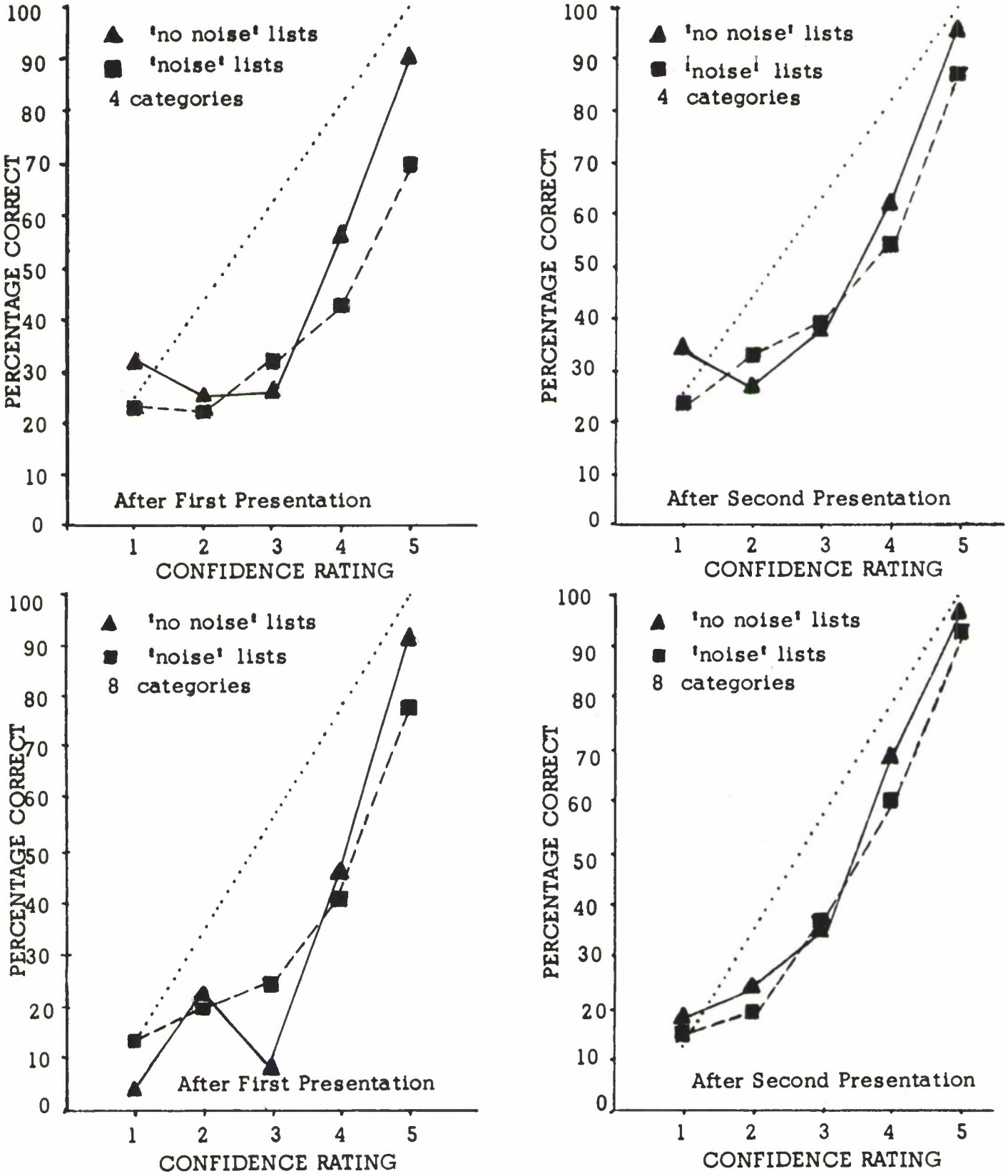


Fig. 7. For Series 3, the correctness-confidcence relations are shown for retention scores following first presentation of acquisition list (on the left) and following second presentation (on the right). Results are shown separately for lists having irrelevant material interspersed ('noise') and for lists without such material ('no noise'). The 4-category sections are shown in the upper half of the figure, while the 8-category sections are shown in the lower half.

'ideal' at the end points and sag characteristically in the middle, showing general overconfidence.

The primary effect of the second-exposure presentation or learning trial was to decrease this 'sag' and to improve the relation of the mid-range confidence ratings to the 'ideal' function line. The improvement in retention or recall was of course significant and was accompanied by a significant increase in mean confidence score.

Summary

The over-all picture from these experiments is as follows:

1. Subjects are consistently able to discriminate their true recollections from mere guesses. Their accuracy is especially notable at the 'anchor points' in the scale: namely, a) when they say their answers are chance guesses, the percentage correct is virtually at the chance level, and, b) when they express 'complete confidence' in their accuracy, their performance score is high.

2. In all comparisons of meaningfulness or association-value of response categories, the material with higher association value was recalled significantly more often than was the lower association value material. In all series the mean confidence scores for high association-value categories were likewise significantly higher than were the mean confidence scores for less familiar response categories. The confidence-correctness relation does not appear to vary over the conditions tested in these experiments.

3. Confidence ratings accurately reflect changes in the learning process--as learning improves, confidence ratings rise also. In addition it appears that confidence ratings tend to improve in realism -- they show a closer agreement with the 'ideal' confidence-accuracy function.

4. Whether the nonsense syllables were associated with 4 categories or with 8, there were no differences in either recall scores or in confidence scores.

Irrelevant material in the form of unassociated nonsense syllables has no clear effect in the confidence-correctness relation.

Obviously it is of considerable importance in many areas of human endeavor if individuals can provide an appropriate and fairly accurate assessment of their own knowledge or memory or perception.

References

Carterette, E., and Cole, M. A Comparison of the Receiver Operating Characteristics for Messages Received by Ear and by Eye. Washington, D. C. Dept. of Navy, Office of Naval Research, 1959. (Tech Rep. 2)

Decker, L., and Pollack, I. Confidence Ratings and Message Reception for Filtered Speech. J. Acoust. Soc. Amer., 1958, 30, 432-434.

Nickerson, R. S., and McGoldrick, C. C. Confidence, Correctness and Difficulty with Non- Psychophysical Comparative Judgments. Percept. Mot. Skills. 1963, 17, 159-167.

Nicol, E. H., Roby, T. B., and Farrell, F. M. Variables Influencing Information Exchange Within Groups. Amer. Psychologist, 1962, 17, 397 (Abstract).

Pollack, I., and Decker, L. R. Confidence Ratings, Message Reception, and the Receiver Operating Characteristic. J. Acoust. Soc. Amer., 1958, 30, 286-292.

Stevens, S. S., Handbook of Experimental Psychology, New York: Wiley, 1951.

Underwood, B. J., and Schulz, R. W., Meaningfulness and Verbal Learning. New York: J. B. Lippincott Co., 1960.

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Decision Sciences Laboratory
Electronic Systems Division
L. G. Hanscom Field, Bedford, Mass. 01730

2a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

2b. GROUP
N/A

3. REPORT TITLE

CONFIDENCE IN RECALL IN PAIRED-ASSOCIATE LEARNING EXPERIMENTS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5. AUTHOR(S) *(First name, middle initial, last name)*
Elizabeth H. Nicol Francis M. Farrell
Thorton B. Roby

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
MAY 1967	21	6

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	ESD-TR-67-457
c. IN-HOUSE REPORT	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	None

10. DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Electronic Systems Division L. G. Hanscom Field, Bedford, Mass. 01730

13. ABSTRACT

The problem of estimating the accuracy of ones own recollections was investigated in four experiments under a variety of conditions. Subjects were shown a series of paired words; then they were shown the first member of the pairs and asked to recall the second member of each. Along with each attempt at recall the subjects were asked to give a confidence rating on a scale from 1 to 5. In all, 180 subjects were tested for a total 11,200 trials. The confidence results are highly significant, indicating that subjects were able to discriminate their correct recollections from mere guesses. Comparisons are presented showing how realism of confidence varies over the main experimental treatments: variations in meaningfulness of material, one versus two exposures to paired-associate lists, and presence of varying amounts of irrelevant material in the acquisition lists.

DD FORM 1 NOV 65 1473

UNCLASSIFIED

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT